AI

OpenAI peels back ChatGPT's safeguards around image creation

Maxwell Zeff 9:13 AM PDT · March 28, 2025

This week, OpenAI launched a new image generator in ChatGPT, which quickly went viral for its ability to create Studio Ghibli-style images. Beyond the pastel illustrations, GPT-4o's native image generator significantly upgrades ChatGPT's capabilities, improving picture editing, text rendering, and spatial representation.

However, one of the most notable changes OpenAI made this week involves its content moderation policies, which now allow ChatGPT to, upon request, generate images depicting public figures, hateful symbols, and racial features.

OpenAI previously rejected these types of prompts for being too controversial or harmful. But now, the company has "evolved" its

approach, according to a blog post published Thursday by OpenAI's model behavior lead, Joanne Jang.

"We're shifting from blanket refusals in sensitive areas to a more precise approach focused on preventing real-world harm," said Jang. "The goal is to embrace humility: recognizing how much we don't know, and positioning ourselves to adapt as we learn."

These adjustments seem to be part of OpenAI's larger plan to effectively "uncensor" ChatGPT. OpenAI announced in February that it's starting to change how it trains AI models, with the ultimate goal of letting ChatGPT handle more requests, offer diverse perspectives, and reduce topics the chatbot refuses to work with.

Under the updated policy, ChatGPT can now generate and modify images of Donald Trump, Elon Musk, and other public figures that OpenAI did not previously allow. Jang says OpenAI doesn't want to be the arbiter of status, choosing who should and shouldn't be allowed to be generated by ChatGPT. Instead, the company is giving users an opt-out option if they don't want ChatGPT depicting them.

In a white paper released Tuesday, OpenAI also said it will allow ChatGPT users to "generate hateful symbols," such as swastikas, in educational or neutral contexts, as long as they don't "clearly praise or endorse extremist agendas."

Moreover, OpenAI is changing how it defines "offensive" content. Jang says ChatGPT used to refuse requests around physical characteristics, such as "make this person's eyes look more Asian" or "make this person heavier." In TechCrunch's testing, we found ChatGPT's new image generator fulfills these types of requests.

Additionally, ChatGPT can now mimic the styles of creative studios — such as Pixar or Studio Ghibli — but still restricts imitating individual living artists' styles. As TechCrunch previously noted, this could rehash an existing debate around the fair use of copyrighted works in AI training

datasets.

It's worth noting that OpenAI is not completely opening the floodgates to misuse. GPT-4o's native image generator still refuses a lot of sensitive queries, and in fact, it has more safeguards around generating images of children than DALL-E 3, ChatGPT's previous AI image generator, according to GPT-4o's white paper.

But OpenAI is relaxing its guardrails in other areas after years of conservative complaints around alleged AI "censorship" from Silicon Valley companies. Google previously faced backlash for Gemini's AI image generator, which created multiracial images for queries such as "U.S. founding fathers" and "German soldiers in WWII," which were obviously inaccurate.

Now, the culture war around AI content moderation may be coming to a head. Earlier this month, Republican Congressman Jim Jordan sent questions to OpenAI, Google, and other tech giants about potential collusion with the Biden administration to censor AI-generated content.

In a previous statement to TechCrunch, OpenAI rejected the idea that its content moderation changes were politically motivated. Rather, the company says the shift reflects a "long-held belief in giving users more control," and OpenAI's technology is just now getting good enough to navigate sensitive subjects.

Regardless of its motivation, it's certainly a good time for OpenAI to be changing its content moderation policies, given the potential for regulatory scrutiny under the Trump administration. Silicon Valley giants like Meta and X have also adopted similar policies, allowing more controversial topics on their platforms.

While OpenAI's new image generator has only created some viral Studio Ghibli memes so far, it's unclear what the broader effects of these policies will be. ChatGPT's recent changes may go over well with the Trump administration, but letting an AI chatbot answer sensitive questions could land OpenAI in hot water soon enough.

OpenAI's viral Studio Ghibli moment highlights AI copyright concerns

ChatGPT's image-generation feature gets an upgrade

Topics: AI AI image generators ChatGPT OpenAI

Maxwell Zeff
Senior Reporter, Consumer

Maxwell Zeff is a senior reporter at TechCrunch specializing in AI and emerging technologies. Previously with Gizmodo, Bloomberg, and MSNBC, Zeff has covered the rise of AI and the...

View Bio >